Distr.
GENERAL

ECE/CES/SEM.54/14
18 May 2006

ENGLISH ONLY

UNITED NATIONS STATISTICAL COMMISSION and ECONOMIC COMMISSION FOR EUROPE CONFERENCE OF EUROPEAN STATISTICIANS

EUROPEAN COMMISSION STATISTICAL OFFICE OF THE EUROPEAN COMMUNITIES (EUROSTAT)

ORGANIZATION FOR ECONOMIC COOPERATION AND DEVELOPMENT (OECD) STATISTICS DIRECTORATE

Joint UNECE/Eurostat/OECD Seminar on the Management of Statistical Information Systems (MSIS)
Sofia, Bulgaria, 21-23 June 2006

Topic (ii): Dissemination and client relations

MULTIDIMENSIONAL STATISTICAL DATA DISSEMINATION ON THE WEB

Supporting Paper prepared by Stefano De Francisci, Giuseppe Sindoni, Istat, Italy and Leonardo Tininini, IASI-CNR, Italy

Summary

I. INTRODUCTION

1. A statistical web warehouse is a system specifically designed for Web-based dissemination of statistical data. Data are stored in a data warehouse (Kimball, 1996) and are accessible from the Web through hypermedia navigation functions, enabling the user to select and dinamically visualise data in various formats. There is a close correspondence between statistical databases and data warehouses as shown in (Shoshani, 1997):

- **microdata** correspond to **fact tables;**
- **macrodata** correspond to **data cubes**
- **macrodata summary attributes** correspond to **cube measures;**
- **category attributes** correspond to **dimensions**;
- **classification hierarchies** of category attributes correspond to **dimension levels;**
- **roll-up** operations provide **summarisation**, i.e. operations to shift from a more to a less detailed aggregation level;
- **drill-down** operations provide shifting from a less to a more detailed aggregation level.

2. As a consequence, one may think that data warehouses can be straightforwardly used for effective Web-based statistical data dissemination. Unfortunately, some peculiarities of statistical data with respect to business data call for extensions to data warehouse techniques with specific models and structures. One peculiarity is about surveys based on samples. Aggregates coming from sample microdata normally require significance checks which are not needed in traditional data warehouses. Another peculiarity is related to privacy and *secondary disclosure*. Organisations responsible for data dissemination must be reasonably sure that disclosure control rules on aggregates are not violated. A last peculiarity is about *sparse tables*, i.e. statistical tables with a high percentage of empty cells, due to the use of meaningless classifications (and/or classification item) combinations. Due to these requirements, traditional data warehouse systems cannot be used for Web-based statistical dissemination without any customisation, because they would allow users to arbitrarily navigate all dimensions available for a given fact, without providing functions to check for conformance to the above principles.

3. This paper presents a generalised technique for Web-based dissemination of statistical data, the underlying conceptual model for spatio-temporal multidimensional data and the provided data warehousing functions. This technique has been used to design and implement a generalised system, called DaWinci/MD (Sindoni, & Tininini, 2006), currently used for data dissemination at Istat.

II. THE DAWINCI/MD CONCEPTUAL MODEL AND NAVIGATION PARADIGM

4. The DaWinci/MD system is based on a model for statistical tables, specifically designed for Web-based data access. The model is based on a decomposition of the information space into *basic multidimensional tables*, represented by a pair of components:

- **object of interest**, i.e. the table measure (e.g. a summary attribute),
- **set of classifications**, i.e. the dimensions used to classify the table measure.

5. Each basic table can have several spatio-temporal instantiations (*spatio-temporal multidimensional tables*), each corresponding to a specific data **time stamp** and **spatial context**.

6. Hence, when choosing a spatio-temporal multidimensional table to be visualised on the Web, each selection step defines incrementally the four table components. Then, a typical query to the system specifies all or part of the components of the <*t*, *s*, *o*, *c*> quadruple, where:

- ***t*** is the time stamp (i.e. *When*);
- ***s*** is the spatial context (i.e. *Where*). It is a combination of:
  - a territorial detail ***d***;
  - a geographical area ***a***;
- ***o*** is the object (i.e. *What*);
- ***c*** is a possibly empty set of classifications (i.e. *How*).

7. The above table decomposition is also the conceptual basis of the information storage model, which represents the spatio-temporal table availability in the system metadatabase. In this

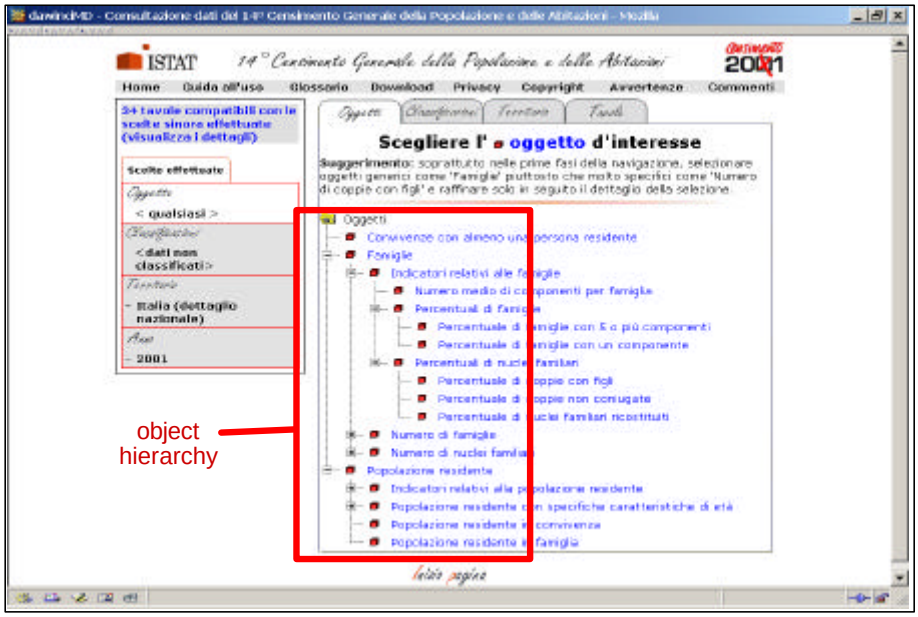
way, the storage of a <*t*, *d*, *o*, *c*> quadruple means that the basic table defined by the <*o*, *c*> pair is available for the *t* time stamp and for all territorial details up to *d*.

8. In the proposed model, data of interest can be chosen according to the above four parameters. According to data warehouse terminology, each parameter quadruple identifies a data *cube*, where a set of measure (the object) values are organised in a multidimensional structure defined by a set of dimensions (the classifications, space and time). The model takes also into account the temporal evolution of territorial hierarchies (Tininini et al., 2002).

A. Object, classification and spatial context selection

9. Statistical objects are organised in hierarchies (see Fig. 1), where a specialisation relationship holds between parent and child objects. In order to facilitate navigation, "virtual" objects can be part of the hierarchy. They group conceptually more specific objects, and they do not correspond to any data warehouse measure. In general, the more generic an object, the higher the number of basic multidimensional tables referring to it.

Figure 1: an example of object hierarchy



10. Classifications are also organised in hierarchies (see Fig. 2), where a specialisation relationship holds between parent and child classifications. As for objects, virtual classifications can be defined and generic, more abstract classifications can be chosen to select a higher number of multidimensional basic tables.

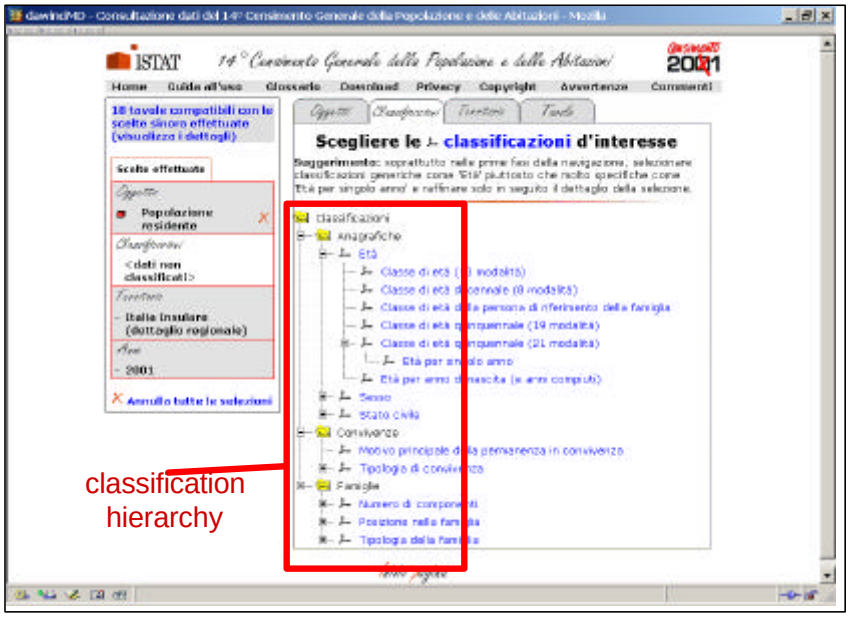


Figure 2: an example of classification hierarchy

11. An object, a set of classifications and a spatio-temporal context identify a set of available multidimensional tables, i.e. a set comprising all tables with the specified, or a more specific, object, the chosen, or more specific, classifications, and such that they can be instantiated in the specified time stamp up to the required territorial detail. A set of selections also determines which further classifications are available for selection by the user (see Fig. 3), namely those classifications forming, together with the previously selected components, a valid <*t*, *s*, *o*, *c*> quadruple.

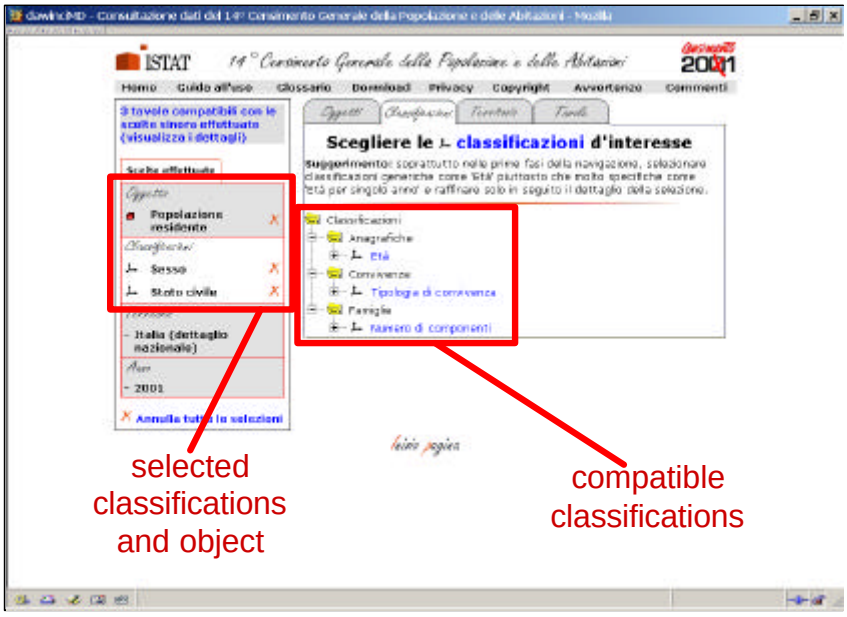


Figure 3: current selections and corresponding compatible classifications

12. Each table is compatible with one or more spatial contexts, depending on the maximum territorial detail, as specified in the system metadata. A geographical area and territorial detail are chosen at the same time (see Fig. 4).

Figure 4: the spatial context selection page

13. They are not independent and the choices available depend on the specific territorial hierarchy. The choice of a higher or lower territorial detail respectively limits or extends the number of tables corresponding to a chosen object and set of classifications.

B. Increasing and decreasing selection criteria

14. Each successive selection of a parameter value decreases the number of further available parameter values and compatible multidimensional tables. The model allows a dissemination manager to publish only the predefined sets of <*t*, *d*, *o*, *c*> quadruples meeting the above confidentiality and significance requirements. In this way users are prevented from requesting tables corresponding to sensitive data or dimension combinations that are meaningless or not planned for dissemination. In this respect, it is possible to choose as many classifications as are allowed by the constraints deriving from the previous choices. It is worth noting that the same generic classification can be chosen more than once, provided that there are tables having at least two classifications both descending from it. Conversely, the removal of a chosen parameter usually decreases the number of constraints and therefore increases the number of further compatible choices.

C. Visualising and navigating multidimensional tables

15. The table chosen from among those compatible with the specified selection criteria is visualised in a Web page (see Fig. 5). Starting from the visualised table, the hyperlinks in the Web page enable the user to remove or add a classification; increase or decrease the territorial or classification detail; change the geographical area. The removal of a classification results in the visualisation of a less detailed statistical table. From the data warehouse point of view, this

corresponds to a roll up operation on the currently visualised cube (table), i.e. to define an *(n-1)*-dimensional sub-cube starting from an *n*-dimensional one. Conversely, adding a classification corresponds to a drill down operation on the visualised cube, i.e. to define an *(n+1)*-dimensional super-cube starting from an *n*-dimensional one that is a more detailed statistical table.

16. A hierarchical classification detail can also be increased (decreased), resulting in a drill down (roll up) operation on the hierarchy. This again results in shifting from a less (more) to a more (less) detailed statistical table. Figure 5 shows the page for the statistical table visualization and warehouse navigation in DaWinci/MD.

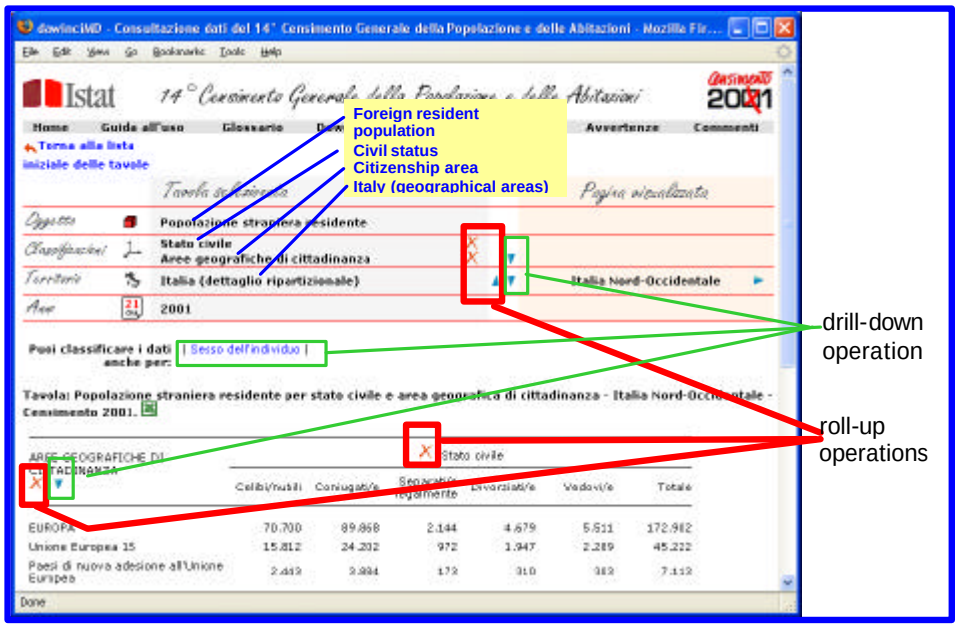


Figure 5: the Web page for statistical table visualization and warehouse navigation

17. The upper part represents a summary of the visualized table, particularly the object (in this case foreign resident population), the territorial area of interest with the chosen territorial classification (in this case Italy and geographical areas), the year of interest and the other classifications (in this case civil status and citizenship area).

18. The statistical table is displayed in the lower part of the page. Roll-up operations can be performed directly by clicking on hyperlinks in the table. Red cross hyperlinks enable the user to remove one classification, while upwards arrows enable the user to navigate along one dimension and to visualize the data at a coarser level of detail. Drill-down operations can also be performed directly by clicking on hyperlinks in the table. For example the user can add a classification (in this case the gender) or increase the detail of an already selected dimension by clicking on the corresponding downwards arrows.

19. Each table can be visualised in one or more Web pages, depending on the number of involved classifications. In Figure 6 a slicing operation is performed by paging the table using the gender classification modalities.

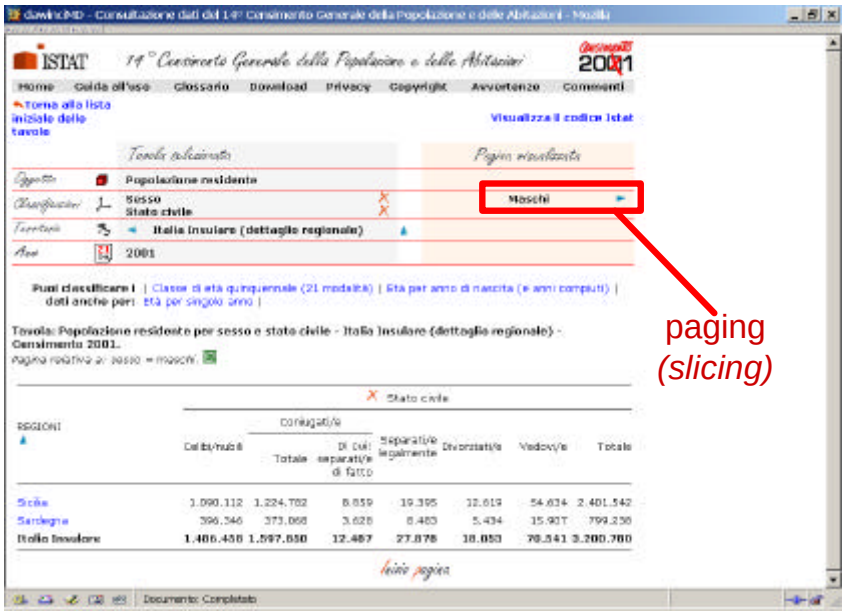


Figure 6: an example of classification paging

20. The territorial detail of the visualised table can also be increased or decreased, thus resulting in a territorial drill down or roll up (see Fig.7). In this case the cube dimensions do not change, but the detail level of one of the hierarchical dimensions defining the cube does change.

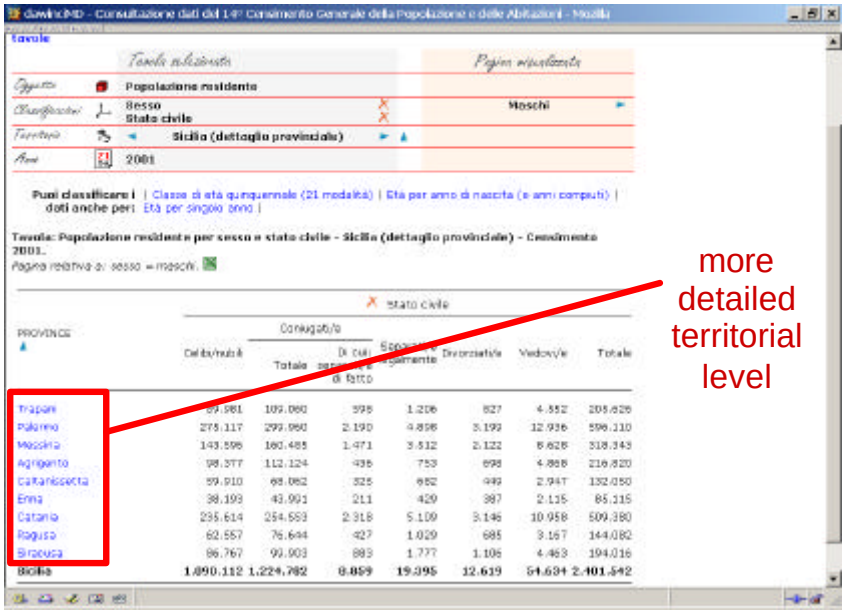


Figure 7: hyperlinks for territorial drill down

III. MAIN RESULTS AND CONCLUSIONS

21. This paper presents the main results reached at Istat on Internet-based statistical data dissemination through an advanced web warehousing system developed in the framework of the Institute's Generalised Dissemination System project (De Francisci et al. 2005). The system

consists of an integrated work platform for online analysis and dissemination of statistical data through multiple layouts, channels and dissemination tools, operating on the various data sources. Its general architecture is designed both to enable designers (representing the class of users enabled for local system management and installation functions) to *package* autonomous system instances, adapting them to their own objectives and choosing the most suitable application solution available for the envisaged purpose, and to give end users the most appropriate tools to interact with the system.

22. In particular, DaWinci/MD is aimed at implementing hypertextual navigation on statistical multidimensional tables and is based on a generic multidimensional model. The model allows a dissemination manager to represent statistical tables containing aggregates at various detail levels in a very effective and flexible way. Navigation paths are driven by the user's current choices, in such a way to provide functions for a free but coherent exploration of the available information space. Exploration is free because the main data warehouse functions in a hypertextual navigation framework are available, and coherent because the system meta-information structure and model ensure both a complete consistency of navigation paths and a full satisfaction of data significance and confidentiality constraints.

23. DaWinci/MD is the component enabling Web dissemination of Istat's Generalised Dissemination System. This is a suite containing a number of products aimed at interactive data analysis and dissemination, all functioning as data warehousing applications but distinguished by various factors:

- *type of data processed:* elementary data, aggregate data, predefined statistical tables;
- *information sources:* validated microdata, other microdata sources, pre-structured statistical tables made available in "local" format by their producers;
- *target environments:* elementary data mart to be accessed by OLAP-functions for interactive analysis; data warehouse of aggregate data or predefined tables explicitly aimed at Web navigation;
- *type of user functions:* aggregation, access and navigation;
- *distribution channels:* intranet and internet for interactive components, off line supports (e.g. CD-ROMs enclosed in the provincial volumes published by the Population Census) (e.g. CD-ROMs to be included in Provincial Population Census book binders)

24. The main aim of the generalised system is to enable the Institute's statistical production areas – and possibly external bodies – to build their own distribution systems, integrated within the production processes as well as with the centralised environments and cross-domain information systems and converging towards integrated output management Institute-wide.

25. Such an installation is based on two principles: *workflow* and *toolkit*[1]. The system runs through *extract*, *transformation* and *load* (ETL) methodologies and technologies, which enable the automated and integrated transformation of validated elementary data into aggregate information for publication. This is then loaded into a generalised database, independently of the statistical domain. This type of process organisation improves the timeliness and coherence of the dissemination process, both minimising the delay between data checks and publication – i.e. the moment that data are returned to the community in a usable form – and, thanks to the high degree of automation, reducing the probability of human error during calculation and storage of aggregate data in the dissemination database.

26. Furthermore, a software module enables Istat to automatically produce a version of the system on a CD-ROM, whose contents are perfectly aligned with the online version and which can be used from any OS platform with a standard software equipment. Special features introduced while developing the databases and optimising software have reduced response times and thus facilitate users with older computers or slow internet connections. The software is structured in such a way that users do not have to install any specific application to access the statistical information: a normal Web browser is sufficient to access all functionalities of the system, which is free of any "invasive" technologies with respect to the user's computer.

27. This system is the culmination of a series of experimental and operational projects conducted at Istat in recent years, aimed at exploring the applicability of state-of-the-art information technologies to the production and dissemination of statistical data. It is the evolution of a number of already-operational systems dedicated to the Web dissemination of data from various surveys, including the Disability Information System, Population Census (from provisional to definitive data), Industry Census (provisional data), law I.S. and Water Census I.S. These systems, which are all available online, were used as prototypes for the generalised information system and are evolving towards a broader-ranging system, which aims at both collection of all *new* dissemination requirements originating from statistics producers and integration with the Institute's numerous existing dissemination systems. The latter, far from being lost, will actually be enhanced by becoming part of a harmonised scenario. This evolution in information system development processes was made possible by the use from the very start of object-oriented design and development technologies, which favour the re-use of design patterns and software modules, thus speeding the development of new public services while keeping costs down.

References

De Francisci S., Renzetti M., Sindoni G., Tininini L. (2005) La modellazione dei processi nel Sistema Informativo Generalizzato di Diffusione dell'ISTAT. *Documenti ISTAT*.

Kimball R. (1996). *The data warehouse toolkit*. John Wiley & Sons.

Shoshani A. (1997). OLAP and Statistical Databases: Similarities and Differences. *Proceedings of the PODS 1997 Conference*.

Sindoni G., Tininini L. (2006) Statistical warehousing on the Web: navigating troubled

[1] The term *toolkit* is commonly used in the computer programming domain to denote a collection of generalised tools and components, which can be used to implement a unified system, customized according to the user's specific needs and requirements. Toolkits are usually made available as libraries or application frameworks

waters. *Proceedings of the International Conference on Internet and Web Applications and Services*. IEEE Computer Society Press.

Tininini L., Paolucci M., Sindoni G., De Francisci S. (2002) Spatio-temporal Information Systems in a Statistical Context. *Proceedings of the 8th International Conference on Extending Database Technology*.
